

Peer-review report of

Pretus, C. S., Gil-Buitrago, H., Cisma, I., Hendricks, R. C., & Lizarazo-Villareal, D. (2024). Scaling crowdsourcing interventions to combat partisan misinformation. *advances.in/psychology*, 2, e85592. <https://doi.org/10.56296/aip00018>

Round 1

Dear Authors,

Thank you for submitting this thought-provoking piece to *advances.in/psychology*. I had the privilege of receiving reviews from two experts in the field of misinformation, and I am grateful for their detailed and constructive feedback. As per standard procedure, I evaluated the manuscript independently before reading their reviews. We all recognize the clear strengths in your work. Your proposals are innovative and emphasize the importance of applying behavioral insights on a large scale. Importantly, it is a clear strength that you outline concrete ways in which to implement these insights. Further, the paper is well-written, and I appreciated your balanced focus on both the strengths and weaknesses of the proposed implementations. Consequently, as an editor, my objective is to enhance the robustness of your work as I perceive considerable potential for it to evolve into a seminal and forward-thinking contribution to the field.

To fully realize its potential, we believe that a series of substantial revisions is necessary. I will outline several key points here, but I request that you address each of our comments in a detailed, itemized revision letter. This letter should also contain excerpts that demonstrate how you have addressed each point, thereby facilitating the subsequent review process.

Reviewer 1 raises a crucial point regarding the necessity to establish the extent to which crowdsourcing fact-checking is effective, beyond the basic assumption that averaging opinions across a group brings us closer to the truth. It is also important to clarify what “fact checking” entails in this context. Mere selections indicating whether a message is true or false may only reflect biases from individuals on both the left and the right, casting doubt on whether this process truly constitutes fact checking or merely acts as a mechanism similar to a like and dislike button. Should the task require participants to conduct their own “research,” it is essential to outline what this process would entail and identify potential sources of error. Similarly, Reviewer 2 inquires about the conditions under which fact checking may not yield accurate judgments. For example, it is vital to consider what sociopolitical conditions are necessary for the assumption of effective crowdsourced fact checking to hold. This consideration is essential to understand the boundaries and limitations of this approach within different contexts and on different social media platforms.

As Reviewers 1 and 2 suggest, I encourage you to further refine and develop your model. Reviewer 2 points out several critical aspects that would benefit from more explicit consideration, such as "hypothesized mechanisms, boundary conditions, falsifiable predictions, assumptions." While some of these aspects are already discussed in the text, organizing them systematically, perhaps through the use of subheadings, spelling them out more explicitly, and providing further elaboration could significantly enhance the impact and clarity of your ideas and their structural coherence.

Relatedly, Reviewer 2 calls for greater precision and a more critical evaluation of the evidence supporting each of the three proposed factors influencing crowdsourced fact-checking. This approach will allow readers to better understand the uncertainties surrounding the proposed roles of these factors. I concur with Reviewer 1 that the section on cognitive dissociation, in particular, could benefit from more precise language and illustrative examples to clarify the concepts discussed.

Reviewer 2 also suggests incorporating a Bayesian perspective on belief updating, which could provide a robust framework for understanding how individuals revise their beliefs in light of new information, although such a process of course needs to be considered in the light of partisanship.

I agree that it seems as if you propose an inverse U-shaped relationship between cognitive dissonance and attitude change. Elaborating on research that has explored this "sweet spot"—where individuals are most receptive to accurate information that contradicts their preexisting views—would be highly beneficial. This discussion could provide valuable insights into the conditions under which crowdsourced fact-checking is most effective.

The current representation in Figure 1 appears to be insufficient in capturing the complexities discussed in your text concerning the factors influencing crowdsourced fact-checking. To enhance the utility and impact of the figure, it would be beneficial to incorporate moderating factors and delineate the interrelationships and potential friction points between different factors. For example, illustrating the tension between achieving cognitive dissonance and maintaining trust in sources would provide a clearer and more nuanced visual representation of the dynamics at play. In short, as you continue to refine and elaborate your model within the text, please similarly develop Figure 1 into a more comprehensive conceptual model. This enhancement will not only improve the visual representation of your theoretical framework but also facilitate a deeper understanding and appreciation of the complexities involved in crowdsourced fact-checking.

Reviewers 1 and 2 have requested additional details on how your approach to crowdsourced fact-checking would be implemented, as well as a more critical evaluation of some of the proposed implementations. Specifically, they inquire about the basis for defining networks within your model. Clarification on how your approach would address issues that do not align with traditional partisan divides is also needed, as both reviewers have highlighted this as a point of concern.

Another crucial aspect of your implementation that requires further exploration is the transparency and signaling of the composition of the crowdsourcing sample to users. While transparency regarding who is participating in the fact-checking process is valuable, it is unlikely that users will investigate the profiles of enough participants to accurately gauge the political composition of the sample. Moreover, the basis on which users would make such assessments is not clear. An alternative you might consider is the use of machine learning models to estimate the political leanings of all users involved in the fact-checking process. These estimates could then be used to weight the fact-checking contributions accordingly and display this information to the recipients of the crowd-checked content. However, this method also introduces potential errors and ethical challenges, such as privacy concerns and the accuracy of the political leaning estimations. Therefore, a thorough discussion of these issues is essential. Explaining how the crowdsourcing mechanism would be perceived and interacted with by end users, and addressing the complexities involved in such a system, would significantly strengthen your manuscript.

Reviewer 2 raises significant inquiries concerning the assumption that central individuals in networks are more accurate or politically moderate. While you acknowledge that future empirical verification is necessary, it would be beneficial to elaborate further on this hypothesis. Specifically, it is crucial to explore the underlying mechanisms that might enable such a pattern, as well as the potential for manipulation by users. Analyzing how centrality could influence political moderation or accuracy, and considering scenarios where these attributes might be exploited, can provide a more comprehensive understanding of the dynamics at play.

You write: “That is, allowing everyone to fact-check but displaying fact-checks from “fact-checkers who are 2 steps away from you” It would be great if you discussed in more detail why you went for 2 or more degrees of relatedness. What factors can be thought to determine the optimal distance in the network?

As a cultural psychologist, I agree with Reviewer 2 on the necessity of being aware of the demographic characteristics of sample populations and the extent to which findings are generalizable. This consideration is particularly critical given that research indicates the left-right political spectrum does not apply universally across non-Western countries.

In the intro, you write: “Efforts to correct misinformation through debunking and accuracy nudges have been found to be ineffective, especially among far-right partisans (DeVerna et al., 2022; Pretus et al., 2023; Rathje et al., 2022) who spread the largest share of online misinformation (Guess et al., 2019).” Please take into consideration this recent paper showing *some* effectiveness even for those at the political extreme. <https://journals.sagepub.com/doi/10.1177/09567976241232905>
Reviewer 2 also asks you to consider a relevant meta-analysis.

Regarding current fact-checking on X: “However, the reliance on single users to generate and validate alternative explanations to misleading posts causes delays in fact-checking, which often occurs only after posts have circulated widely and accumulated thousands of impressions.” A bit more precision would be good as the validation part indeed seems to be based on large groups of users (i.e., crowdsourcing).

“This is because of a phenomenon known as *wisdom of the crowds*, which was first discussed almost a century ago (Galton, 1907),” Please change to “more than a century ago”.

In Figure 2, in the left panel, please also highlight the fact-check receiver as you do in the right panel.

On p. 20, you write “Extreme partisans may be less willing to fact-check ingroup sources”. However, from a partisan viewpoint, it is equally plausible that “fact-checking” could serve to reinforce the credibility of messages that align with ingroup beliefs, thereby making it attractive for partisans to engage in fact-checking. This alternative perspective suggests that partisans might actively participate in fact-checking to enhance the persuasive impact of ingroup-consistent information.

For pedagogic purposes, please create a table that summarizes the two proposed implementations, including their strengths and limitations. The table could be combined with Figure 2.

Style:

Please be consistent in the use of round or straight ‘

For high resolution, please provide figures as SVG files (inserted into the document).

The reference list needs a lot of attention:

1. Please make sure all article titles in the reference section are sentence capitalizes (instead of capitalizing each word) in line with APA 7
2. Please also make sure that all journals are word capitalized (except for words under 4 letters) in line with APA 7
3. Please make sure that all DOIs in the references are presented and, importantly, as URLs. Also, some DOIs are presented twice or in the wrong location.
4. Please avoid full uppercase for titles, e.g., "NBER WORKING PAPER SERIES A MODEL OF ONLINE MISINFORMATION A"
5. Some journal names are missing.
6. Some publishers of books or reports seem to be missing. For reports, URLs always have to be presented.

There is a small typo on p. 20: "Knowing how much each group is willing to fact-check different political sources and how responsive it is to different crowd sizes would help fine-tune whose ratings should users **should** have access to."

and on p. 21: "Elucidating these questions will be key to confirming the suitability of network-based approaches to define whose ratings are people **are** exposed to."

On p. 12, do you mean "broad" instead of "board"?

In conclusion, we recognize significant potential in this innovative contribution and aim to enhance its robustness and comprehensiveness. By addressing the concerns raised in the evaluation by the Reviewers and me, you should be well-positioned for improvement. Please note that the word limit for reviews is 10,000, excluding references, abstracts, figures, and tables, which should provide ample space for your revisions. In light of the extensive revisions required, I would appreciate receiving a revised manuscript within two months. Therefore, please disregard the automatic reminder emails set for the standard one-month period. Of course, you can submit the revised manuscript earlier if you prefer.

Thank you once again for submitting your intriguing work to our special issue. I look forward to receiving the revised version.

Best,

Jonas R. Kunst

Editor-in-Chief

Reviewer 1:

My thanks to the authors and editor for the opportunity to review this submission. It presents a review arguing that three factors are likely to affect the efficacy of crowdsourced fact-checking interventions on social media (cognitive dissonance, trust, and crowd size), provides recommendations of how these factors can be considered, and specific tactics for fulfilling these recommendations based on ideas from network analysis.

I agree that misinformation is an important danger, and I really liked the authors' attempt to apply psychological knowledge to actually *do* something about. I also liked the boldness of some of the ideas; this certainly wasn't the average incremental empirical paper, but something much more creative. I nevertheless did pick up a range of issues (some major, some minor) that I have elaborated upon below.

MAJOR POINTS

1. The abstract and page 12 describe the paper as presenting a “model”. Obviously, there's a degree of subjectivity involved in deciding constitutes a model, but I don't personally think this term is an accurate description of what is presented in this manuscript. There are very few of the elements I'd expect to see in an article proposing a theory or model – i.e., clear identification of the phenomena to be explained by the model, mathematical or computational formalization, hypothesized mechanisms, boundary conditions, falsifiable predictions, assumptions, etc. To me, this manuscript is better described as a narrative literature review or essay. The claims about cognitive dissonance, trust and crowd size seem more just like three separate hypotheses. If the authors do wish to engage in theorizing, useful guides can be found in Borsboom et al. (2021) and Guest and Martin (2021).
2. In general, I would like to see much more critical discussion about the extent to which evidence supports (or does not support) the claims made in the manuscript. This applies throughout – to the claims about cognitive dissonance, trust and crowd size, to the general recommendations that follow about fact-checking, to the specific mechanisms you propose for enacting those recommendations. I know the paper includes plenty of citations, which is great, but this isn't really enough on its own. It's important also to critically examine the evidence itself, and transparently convey the strength of evidence the reader. As it stands, a reader (say, a decision-maker at a social media company) could easily come away from this manuscript with the impression that it provides solid evidence-backed recommendations, when that isn't really the case – instead it provides ideas that are bold and interesting, but largely untested, and based on quite tentative knowledge about the three main factors under discussion. Bold and untested ideas *are ok*, but the claims you make need to be *calibrated* to the strength of the available evidence (see Hoekstra & Vazire, 2021).
3. I think it would be easier to achieve point 2 if the manuscript was somewhat more focused on covering a limited set of key issues in a logical flowing order.

Currently the manuscript covers a very wide range of ideas, and while I could understand *what* each individual sentence said easily enough in isolation, I often found myself struggling to see how the ideas were connected. E.g., there's a paragraph about homophily in the middle of the cognitive dissonance section, and it doesn't mention the word dissonance. Why is it in there? As a reflection of the same issue, I found it a bit challenging to work out which points were key assumptions of the recommendations you were making, and which were non-crucial background. Obviously, you as the authors know this topic area really well, which is awesome, but you don't need to reel off everything you know about it – it's ok to try to cover just a small quantity of core ideas, but to cover them very critically and comprehensively.

4. Near the end of the manuscript, there is this key recommendation “we propose limiting whose fact-checks are visible to each user based on the relative or absolute location of the fact-checker in social networks.” I know that sometimes we as academics can get a bit stuck in complexities and nuances, so I just want to prompt some commonsense reflection here: Do you *really* think that *reducing* the number of fact-checkers (and consequently the number of fact-checks) people are exposed to will result in people having more accurate beliefs? I'm not saying it's impossible this could work – maybe it could – but it is a rather out-there idea that seems to me to conflict with some fundamental things we know about human beings (e.g., that they do update their beliefs in response to new factual information, albeit sometimes rather slowly; that more people are generally more wise than fewer people; etc.). That all said, perhaps I'm being uncharitable in quoting this sentence in isolation, and you see the picture differently. If so, feel free to elaborate.

MINOR POINTS

1. Abstract: It's a bit unclear what “at odds” means in this context
2. The claim that misinformation is the “most severe short-term risk the world faces” was widely criticised when the WEF released this report (c.f. climate change, wars in Gaza and Ukraine, etc.). I think it's ok to include, but only if it's presented in conjunction with information about how the finding was produced (just a self-report survey), and ideally some degree of critical examination.
3. “Efforts to correct misinformation through debunking and accuracy nudges have been found to be ineffective, especially among far-right partisans”. I don't think this is a fair reflection of the evidence. There are many published examples of debunking interventions showing some corrective effect, even if debunkings are by no means a complete “cure”. See for example the meta-analysis by Chan et al. (2017).
4. Given the importance of the idea of *crowdsourcing* in this paper, it'd be useful to provide a definition early on.
5. “Accuracy then originates from the statistical principle that, while individual estimates may vary widely, their average is likely to be closer to the truth due to the dissolution of individual biases and errors.” I don't think this should be labelled as a statistical principle unless it's possible to show a statistical proof that it's true.

6. The essay includes lots of (perhaps justifiable) enthusiasm about the wisdom of the crowds, but in what situations do crowds *not* produce accurate judgments? (This isn't my area but there must surely be some!)
7. Re. the discussion of cognitive dissonance, I noticed this was framed in terms of how it related to "belief updating". Something to consider here is that, in a Bayesian sense, the *greater* the conflict between new information and existing beliefs, the *greater* the capacity for updating. In contrast, if new information is fully consistent with existing beliefs, there is no "room" for updating. Consequently, isn't it more plausible to expect a positive relationship between dissonance and updating in most situations? (Note: This doesn't necessarily imply a positive relationship between dissonance and the accuracy of the updated beliefs!)
8. "Excessively dissonant opinions can put a person at risk for cognitive immunization, while not enough dissonance can foster the reinforcement of false beliefs shared by ideological communities. Thus, the composition of the crowd should be selected in a way that promotes an optimal level of cognitive dissonance." Here you are hypothesising a non-linear (reverse-U shaped) effect of dissonance on beliefs (or belief updating), and making an action recommendation on that basis. But what evidence is there that such a non-linear effect exists? There's presently no evidence presented in the manuscript that I can see that supports this idea, or that could tell the reader what the optimal level of dissonance is. I also struggled somewhat to grasp what the theoretical basis for this non-linear relationship.
9. There is an uncritical reference to the idea of a backfire effect on page 12. Please acknowledge the evidence of contradictory evidence against this idea (see Wood & Porter, 2019). Even Nyhan himself is now quite skeptical of backfire effects (see Nyhan, 2021).
10. "Crowdsourcing interventions can help overcome this limitation by offering corrections from a board and a heterogeneous group of fact-checkers that do not share a single identity". Doesn't this contradict with the earlier sentence "Knowing that outgroup sources have participated in correcting in-group misinformation could undermine fact-checking efforts"?
11. In the section on Crowd Size, the only evidence that seems to be presented to support the arguments is that "individuals were less likely to share posts that were flagged as misleading by a crowd that was 80% (compared to 20%) of the size of the crowd who had *liked* them (Pretus et al., 2024)." But I don't think that finding tells us much about crowd size – it just tells us that people (naturally) give less credence to claims that have a lower percentage of positive evaluations. It's a bit like saying that people think restaurants with more positive reviews are likely to be better.... Consequently it's not clear whether there's any evidence to support the ideas about crowd size presented.
12. Re. the suggestion of using central people as fact-checkers: I understand the idea here, but is there any reason to believe that central people will tend to have accurate beliefs to share in fact-checks? Furthermore, if any social media company or algorithm gave fact-checking status to people who meet some known definition of

centrality, what would stop partisan groups from deliberately satisfying these criteria to get fact-checking status? It's not that hard to add some new friends or followers, after all...

13. Likewise for people who are relative distances from those needing fact-checking; don't we need to consider whether these people hold *accurate* beliefs?

14. In general there is a lot of focus on *partisanship* and extremes of the political spectrum in the manuscript, and much less grappling with the question of who is more or less likely to hold correct beliefs and thus be more or less suitable as a fact checker. Not all misinformation is politically valenced or partisan. (Or is the paper meant to apply exclusively to fact-checking of political claims?)

15. "center-left and center-right voters are less susceptible to misinformation (Guess et al., 2019) and more politically knowledgeable than people at the extremes and center of the political spectrum (De keersmaecker et al., 2024)." I think this needs more qualification with respect to the countries and time periods this conclusion is based on. Please try not to present observations about specific places, times and political landscapes as if they were human universals.

16. I found the directions for future research interesting, but to me they leapt too far ahead into nuanced questions that assume we know the main claims of the manuscript are true and that its main suggestions for interventions actually work. The manuscript doesn't provide particularly strong evidence to support central contentions like a non-linear effect of dissonance on updating, or that crowdsizes has a positive effect on updating, or that the tactics proposed would be efficacious. To me, these are the most important future research directions; more nuanced research questions like those proposed in the current future directions section are probably best considered further down the track.

REFERENCES

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, 1745691620969647. <https://doi.org/10.1177/1745691620969647>

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11), 1531–1546. <https://doi.org/10.1177/0956797617714579>

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4), 789–802. <https://doi.org/10.1177/1745691620970585>

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, 5(12), 1602–1607. <https://doi.org/10.1038/s41562-021-01203-8>

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15), e1912440117. <https://doi.org/10.1073/pnas.1912440117>

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1), 135–163.
<https://doi.org/10.1007/s11109-018-9443-y>

Reviewer 2:

This manuscript argues that crowdsourcing interventions can be an effective way of addressing misinformation. In doing so, the authors review three factors—cognitive dissonance, trust in sources, and crowd size—that may impact belief update. The authors then contrast the potential of a “centralist” approach and a “two-steps-away” approach to optimizing tradeoffs between the aforementioned three factors during intervention implementation. Finally, the authors discuss some future directions.

Overall, I enjoyed reading the manuscript and found the proposed approaches to be both novel and insightful. With the popularity of social media, much of our communication is now online, and the networked approach to fact-checking has real potential to improve the state of the information ecosystem. I do, however, have several concerns that I think should be addressed or at least acknowledged. I outline these concerns below, broadly organized in the order in which the relevant paragraphs appear in the manuscript.

First, I think we must establish that crowdsourcing can appropriately identify and fact-check misinformation, before we can consider ways to optimize belief update for individuals exposed to those fact-checks. Although the manuscript touches on “wisdom of the crowd”, the discussion around this crucial point is brief and relies only on the general idea that adding diverse viewpoints will lead to accuracy, because of “the statistical principle that, while individual estimates may vary widely, their average is likely to be closer to the truth due to the dissolution of individual biases and errors”. This is not wrong, but requires stringent assumptions that may be more or less tenable depending on the particular network under question. For example, the classic case of Condorcet’s jury theorem requires independence between individuals’ judgements, for accuracy to increase along group size. Although more recent work has attempted to relax such assumptions, my understanding is that there can be a threshold of dependence that, if exceeded, the aggregation of ratings/votes can no longer be expected to be more accurate than the individual (e.g., Hahn et al., 2019). I therefore do not know how likely it is that “wisdom of the crowd” will hold in practice, particularly for the types of social networks that most require interventions. To be clear, however, I do not expect the authors to work out or formally model this within this short conceptual piece. What I do think is that there is a need to explicate further such underlying assumptions. Ideally, this will be integrated/considered in the theoretical model, but, if not, then at least as part of the discussion on future directions.

Second, the section on cognitive dissonance is somewhat unclear. As the authors note at the outset, cognitive dissonance is typically defined as “the underlying tension created by inconsistencies between an individual's behavior and their thoughts and beliefs”. I agree that such cognitive dissonance or discomfort can impact belief updating and misinformation processing (e.g., Susmann et al., 2022). However, I am not sure what are the beliefs and behaviors within the context of the authors' discussion. Is the act of staying within a network of plausibly misinformed individuals the behavior? Is it the choice of not seeking out further information when someone comes across a fact check correcting in-group information? What about the beliefs? From reading, the authors seems to be referring more to a general sense of “cognitive discomfort”. Likewise, the authors also attempts to introduce the notions of cognitive immunization and prediction error in this section, but do not currently provide sufficient explanation to integrate them within the discussion. In my view, including a few illustrative examples would really help readers understand the psychological processes that the authors are claiming may occur.

Third, I think the theoretical model can be better developed and/or more comprehensive. This is not to say that the three factors that the authors identified are unimportant. My concern is simply that, throughout most of the manuscript and particularly in Figure 1's presentation of the model, these factors are assumed to be independent. That is, the edges/arrows point only towards belief update but not each other nor other relevant factors, and there is no discussion on potential feedback loops with belief updating. This is fine as is, if the authors' intention is to list three relevant factors. If so, I suggest removing Figure 1 and/or toning down claims of a novel model. On the other hand, I think going beyond listing three factors and making clearer to readers why those three factors (and not others) are uniquely relevant and/or the potential interactions between them would substantially improve the manuscript.

Finally, I wonder how the interventions will play out in the real world where there are different types of misinformation as well as non-political group identities. The manuscript focuses on partisan misinformation, which is fine and indeed an important part of the information ecosystem. However, social-media networks in the real world are global in nature and can be a complex web of conflicting identities and information, although it depends again on the network in question (e.g., Facebook vs. X/Twitter). Relatedly, and maybe I am missing something, it is also not too clear to me how the network is going to be defined during implementation. Will it be based on interaction, engagement, or friends and follower/following? Can different social-media platforms and different ways of constructing networks affect the proposed interventions? How might dynamic vs. static networks come into play (e.g., individuals in the real world can break ties, potentially if discomfort/dissonance is too high)? Might there be cases in which the interventions backfire and/or increase polarization (e.g., Bail et al., 2018)? How about the characteristics/features of particular network (e.g., algorithms, up vs down votes on Reddit, likes on X, etc.)? The authors will not need to answer all these questions, but I think readers will appreciate a more thorough discussion about the practical difficulties in implementation, at least in the section on future directions.

Minor

- p.1 Abstract mentions AI but AI isn't discussed in the manuscript
- p.2 Polarization can arise even amongst normatively rational agents (e.g., Cook & Lewandowsky, 2016)
- p.8 Ecker et al. (2022) highlighted that belief updating requires the detection of a conflict between misinformation and correction but it did not discuss cognitive dissonance
- p.12 "board" should be "broad"
- p.19 Might be interesting to briefly discuss open-source platforms that might be the implementation of such interventions more feasible (although I suppose then there's the issue of crowd size)

References

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216-9221.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in cognitive science*, 8(1), 160-179.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in cognitive science*, 11(1), 194-206.

Susmann, M. W., & Wegener, D. T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory & Cognition*, 50(2), 435-448.

Round 2

Dear Authors,

Thank you for submitting your significantly improved manuscript. Both the reviewers and I appreciate your attentiveness to our previous comments. The work has become more nuanced and elaborate as a result. However, both reviewers still have some issues they would like you to address. Having read the revised manuscript alongside their comments, I largely concur with their observations and thank them for another round of thorough work. I will not reiterate each of their points as they have explained them succinctly. Once you have addressed these remaining issues, I believe we are nearing the completion of the review process for this work. My intention is to make the next decision without consulting the Reviewers again, but this will depend on your responsiveness to their points.

Thank you for revising your reference list. There are some remaining issues which I am detailing in a separate email for you to address. The writing is generally very good, though there are a few minor typos I noticed. For instance,

There seems to be a typo p. 28: “One of these assumptions IS the potential non-linear effect of dissonance on belief updating:”

On p 29: “Testing THESE two assumptions is a critical first step to validate our proposed model and implementations.”

On p. 31: “Elucidating these questions will be key to confirming the suitability of network-based approaches to define whose ratings people exposed ARE to.”

Thank you once again for revising your paper for our special issue. I believe we are making substantial progress and look forward to receiving a further revised manuscript, preferably by July 28. Please reach out to me if you require additional time.

Best,

Jonas R. Kunst

Reviewer 1:

Thank you for the opportunity to review this manuscript. After reading the revision, I can see that the manuscript has substantially improved, and I want to again commend the authors on the novelty of their proposed approaches. However, I still have several concerns and suggestions. I have also read through the other reviewer's comments and largely concur with their assessment, but I will aim to avoid overlap in our comments. I hope my feedback will be useful anyhow.

1. As a general comment, the manuscript includes many interesting ideas, but the organization and integration of key points can still be improved. For example, while the manuscript now notes that the accuracy of crowdsourcing depends on independent judgments and the method of aggregation, it does not revisit these assumptions or discuss them in relation to either the centralist or two-step away approach. Although the focus of the manuscript is on “the ability of crowdsourced fact-checks to influence belief updates,” it seems to me that these two approaches could also be considered different aggregation methods, impacting the accuracy of the aggregated judgment. Perhaps I am missing something, but it is not clear whether we should simply assume that aggregation is accurate and focus solely on

maximizing the reach and impact of the aggregated judgment. Table 1, which outlines some hypotheses and limitations of the two approaches, is helpful, but it could perhaps benefit from further expansion to reference some of the important factors mentioned at the outset. If not expanded, this discussion could also be integrated into the text, as the authors have done with other assumptions in the future directions section.

2. Regarding the section on cognitive dissonance, I appreciate the added examples. However, in my view, it remains somewhat unclear. The cited definition of cognitive dissonance is “the underlying tension created by inconsistencies between an individual's behavior and their thoughts and beliefs (Festinger, 1957).” Under the cited definition, unless one has already shared misinformation and then come across the fact check (and then updated their beliefs in light of the fact check), it doesn't seem like there would be any inconsistency between behaviors and beliefs. That is, the person's misinformation-sharing behavior would still be in line with their misinformed beliefs, unless they manage to successfully update their beliefs in the first instance. The discussion that follows thus seems to rely on a broader definition of cognitive dissonance that includes tension or discomfort arising not only from inconsistencies between behaviors and beliefs but also between conflicting beliefs. I suggest the authors clarify this and cite the broader definition.

3. The discussion on prediction errors is introduced as “in contrast to cognitive dissonance,” even though it is placed within the cognitive dissonance subsection. This transition seems somewhat abrupt. Additionally, since prediction error is not mentioned again in subsequent sections, its inclusion here feels somewhat disjointed. If the authors wish to retain the point on prediction errors, it would again be beneficial to better integrate it with the rest of the manuscript. Otherwise, it may be more coherent to omit this discussion entirely.

4. The manuscript includes sections on “Why do people believe crowdsourced judgments?” and “When do crowdsourcing interventions influence people's beliefs?”. These sections address very similar overarching questions, which makes their separation seem to me somewhat redundant. Integrating and combining these sections could perhaps improve the flow.

5. The manuscript defines a conversation network as a static network. To my knowledge, static networks are those in which no one can form new ties or break existing ones. However, in real-world social media, networks tend to be dynamic, as people can (and do) easily decide to “follow” or “unfollow” others. The manuscript could perhaps benefit from a discussion on static (as is commonly done in experimental research) versus dynamic (real-world) networks and point to this as another future direction.

6. The authors discuss the use of machine learning in the section on the two approaches. It might be better to move this discussion to the future directions section.

Reviewer 2:

Thank you for your comprehensive and thoughtful response letter, and for the careful revisions you've made to the manuscript. I believe the changes you've made in response to my earlier suggestions largely address my initial concerns, and the manuscript seems much improved. At this stage I just have a few minor remaining suggestions, each of which should be reasonably quick and easy to address. Well done on all your hard work on this paper!

1. "In terms of crowd size, several studies report high levels of external agreement with expert assessments based on the evaluations of 10 crowd workers (Roitero et al., 2018; Roitero, et al., 2023; La Barbera et al., 2020). Espina Mairal et al. (2024) also find that more than 3 or 4 individuals significantly boost the accuracy of trustworthiness estimates. Thus, larger crowds are more likely to offer accurate collective estimates". Here the first sentence doesn't seem to pertain to the effect of increasing crowd size, and thus doesn't really serve as evidence for the conclusion in the final sentence of the paragraph.
2. "Some agencies use four to seven-level scales, such as True, Partly True, Misleading, and False (Lee et al., 2023), while some studies have experimented with 0 to 100 and 0 to infinite scales (Roitero et al., 2018). Recent work suggests that greater agreement with expert judgments can be achieved by merging these ratings into 2 or 3 categories, such as True, In between, and False." Is this comparison literally based on percent agreement? If so, agreement will be artefactually higher with fewer categories.
3. "For instance, people are less willing to share misinformation with every additional person participating in collective accuracy judgments (Pretus et al., 2024), highlighting the impact of the absolute size of the fact-checking crowd". Is this when holding the proportions of negative ratings constant? From a quick scan, I struggle to find an analysis in Pretus et al. that specifically looks at the effect of absolute crowd size while holding other features constant.
4. "Effective interventions against misinformation should generate enough cognitive dissonance to facilitate belief update, without inducing cognitive immunization". This touches on points in my previous review. I'm afraid there still isn't quite enough evidence presented to justify this action recommendation. We first need to establish that the effect of cognitive dissonance on belief accuracy (or belief updating) is of the non-linear shape assumed here. You could rephrase this sentence to make it clear it rests on an untested assumption, or leave this action recommendation until you discuss it in more detail later.
5. "Moreover, to avoid backfire effects, collective accuracy judgments should come from trusted sources in different communities". I think it would be useful to re-frame this. There is little evidence that backfire effects exist at all. But your case for considering trust (as presented elsewhere in the paper) doesn't seem to hinge solely on trying to avoid backfire effects.

6. An interesting feature of the model in Figure 1 is that it implicitly assumes that shared group membership only affects crowd size via cognitive dissonance and trust in sources (i.e., a completely mediated effect). Is this deliberate, or is there just an arrow that needs to be added? Likewise, the model assumes that any effect of shared group membership on belief updating is fully mediated by the variables in the middle of the figure. Is this deliberate?
7. "One way to overcome this would be to use fact-checking exclusively as a dissuasive tool rather than a persuasive tool. This could be achieved by allowing fact-checkers to "downvote" content by marking it as "misleading" but not "upvote" content by marking it as "credible"". Wouldn't there be downsides to this? I could imagine a case where a claim is accurate, the vast majority of fact-checkers perceive the claim as accurate, yet the crowd size is sufficient for the claim to receive many negative ratings....